DOCUMENT RESUME

ED 123 257 TM 005 316

AUTHOR Taylor, Terry D.; And Others
TITLE Measurement of Classroom Versus Written Exam Teacher Performance; Should Teachers Show or Tell?
INSTITUTION Far West Lab. for Educational Research and Development, San Francisco, Calif.
PUB DATE [Apr 76]
NOTE 29p.; Paper presented at the Annual Meeting of the American Educational Research Association (60th, San Francisco, California, April 19-23, 1976)

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS Comparative Analysis; Early Childhood Education; *Educational Background; Effective Teaching; Inservice Teacher Education; *Measurement Techniques; Observation; *Preschool Teachers; Teacher Characteristics; Teacher Education; *Teacher Evaluation; Teaching Skills; Testing; Test Reliability; Test Validity; *Validity
IDENTIFIERS *Flexible Learning System

ABSTRACT

Paper-pencil tests and classroom work-samples were used to evaluate the effectiveness of three learning units for teachers in early childhood education. Half of the sample had BA degrees or above, and half high school diplomas or less. Both methods assessed the same four competencies: quality of classroom arrangement, personalization of a literature experience, awareness of sense perception uses of materials, and diversity of teaching strategies which develop children's sense perception. Stability coefficients for the paper-pencil measures ranged from .29 to .83; generalizability coefficients for the work-sample measures ranged from .67 to .98. In general, validity coefficients increased from pre to post for the high school educated teachers, but decreased for the college educated teachers. The validity coefficients were higher than the average correlations between different competencies assessed by different methods. This suggested some degree of discriminant validity of the competency measures. However, the mean correlations between different competencies assessed by the same method suggested that at pre-test, the paper-pencil test might be the more appropriate method for the college educated teachers, and the work-sample method might be more appropriate for the high school educated teachers. After training, discriminant validity improved for both groups. (Author)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED123257

Measurement of Classroom Versus Written Exam Teacher Performance: Should Teachers Show or Tell?

Terry D. Taylor, Stanley H. L. Chow,
Judith J. Hubner, and Steven P. Stripp

Far West Laboratory for Educational Research and Development
San Francisco, California

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

TM005 316

Abstract

Paper-pencil tests and classroom work-samples were used to evaluate the effectiveness of three learning units for teachers in early childhood education. Half of the sample had BA degrees or above, and half had high school diplomas or less. Both methods assessed the same four competencies: Quality of classroom arrangement, Personalization of a literature experience, Awareness of sense perception uses of materials, and Diversity of teaching strategies which develop children's sense perception. Stability coefficients for the paper-pencil measures ranged from .29 to .83; generalizability coefficients for the work-sample measures ranged from .67 to .98. In general, validity coefficients increased from pre to post for the high school educated teachers, but decreased for the college educated teachers. The validity coefficients were higher than the average correlations between different competencies assessed by different methods. This suggested some degree of discriminant validity of the competency measures. However, the mean correlations between different competencies assessed by the same method suggested that at pre-test, the paper-pencil test might be the more appropriate method for the college educated teachers, and the work-sample method might be more appropriate for the high school educated teachers. After training, discriminant validity improved for both groups.

Historically, the evaluation of student and in-service teachers has involved use of paper-pencil examinations, clinical judgments of efficacy by supervisors or master teachers, or classroom observation of teaching behaviors. An implicit assumption underlying these methods of assessing teacher performance is that they can meaningfully measure the knowledge, traits, attributes, or competencies they are intended to measure among populations with diverse background characteristics.

There is recent evidence, however, that challenges this assumption. Taylor, Marx, and Winne (1975), for example, found that methods of measuring children's self-concept may interact with particular characteristics of the populations they are used with. They concluded that "self-concept instruments purporting to measure similar constructs may be measuring different constructs due to variations in the features of norming samples." Stretching this conclusion to its limit, one might infer that an instrument is measuring what it is intended to measure only within the situation in which it was developed and normed. Thus, the generalizability of interpretations resulting from educational research rests on shaky foundations until the variables and processes associated with measurement methods are systematically studied.

Testing the validity of using teacher assessment methods with diverse populations became the objective for this study based on findings from prior research on teacher performance. Taylor and Chow (1974) found that preschool teachers with a relatively extensive educational background (e.g., B.A. degrees or teaching certificates) tended to perform better on paper-pencil exams assessing knowledge, understanding, and application of particular areas of teaching competency than their less well-educated colleagues (high school diploma or less) before and after participation in an educational training

course. They suggested that the differences may have been due to differences in learning ability, but could just as well have resulted from differences on test-related variables such as test taking ability, reading comprehension, or writing skills.

The validity of teacher evaluation methods is of vital importance in the field of early childhood education. Preschools in many states do not require extensive formal teaching preparation and consequently teachers tend to vary widely in their educational background. Given this diversity, can one expect the process of responding to a paper-pencil exam to be as appropriately geared to teachers with limited education as it is to teachers with extensive familiarity with the formal educational process? Might not another method such as classroom behavior measurement have greater validity for teachers without extended education?

These questions became the focus of the study reported in this paper. Specifically, we wished to investigate the validity of two methods of measuring several teaching competencies within two groups of teachers varying in formal education. Convergent and discriminant validity of the two methods -- a paper-pencil exam and a classroom performance method -- were investigated using the multitrait-multimethod matrix proposed by Campbell and Fiske in 1959. Essentially, the matrix allows examination of three facets of validity: a) convergence of independent measurement procedures on a particular psychological trait (or, in this case, teaching competency), b) discriminant validation of measurement procedures constructed to measure a trait or competency from methods measuring other traits or competencies, c) estimates of variance due to method of measurement procedures rather than representative of the trait or competency intended to be measured.

The study is a part of a larger research effort to investigate the effectiveness of an early childhood teacher education program -- the Flexible Learning System (FLS) -- developed at Far West Laboratory in San Francisco. The FLS is a process-based program designed to develop teaching competencies valuable in the early childhood classroom (preschools through third grade). The larger study examined the effects of training with three FLS teacher education modules on teacher performance.

Method

Subjects

A subsample of 32 teachers was drawn from a sample of 110 in-service preschool teachers and assistants employed in Head Start or Children's Centers Programs in San Diego, California. All participants had the autonomy as well as the opportunity to implement changes in classroom arrangment if they so desired. Due to the attrition of one teacher, the final sample of the study consisted of 31 teachers and assistants.

The 32 teachers were selected on the basis of academic degree attained: 16 participants had high school diplomas and 16 participants had B.A. degrees. Twenty-seven participants were teachers and four were teaching assistants. Eighteen of the teachers were under 30 years old; eight were between 30-45 years old; four were over 45 years old. The sample consisted of mostly female teachers (n=27) who were working with preschool students. About half of the sample (n=15) had teaching credentials; less than half of the sample (n=12) were parents. Almost all of the teachers had one to eight years of teaching experience: six had one year experience, thirteen had three years or less experience, and nine had eight years or less experience.

The sixteen "high education" teachers, as a whole, tended to be younger (n=13 under 30) than the "low education" teachers (n=5 under 30), had no children, and possessed teaching credentials. They tended to have slightly less working experience with children than the teachers with lower levels of education. For example, only four of the 16 high education teachers had eight years experience or more as compared to seven of the teachers with low education. On other variables, such as teaching position, sex, and grade level of the children they worked with, the groups were approximately equivalent.

Procedures

As mentioned earlier, 32 teachers were blocked on educational level (16 teachers with a BA degree or above, 16 with a high school diploma or less). Teachers within the two groups were then randomly assigned to one of the three treatments or a control condition, resulting in four high education and four low education teachers per condition. Each of the three treatment groups received training in one of three FLS teacher education modules.

The three modules selected were (a) Physical Arrangement of the Classroom, (b) Enriching Literature Experiences of Children, and (c) Developing Children's Sense Perception.

The training was delivered in one 2-3 hour session each week for a total of six weeks by an instructor trained by Far West Laboratory staff. Three weeks before training began, all teachers received a set of instructions explaining that they would be observed twice before training and twice after, and indicating how they were to prepare for these observations.

At the same time, each teacher received a set of instructional materials for preparing a sensory activity and a reading activity to be taught on the observation occasions. Data on teacher performance during these activities as well as the physical arrangement of the classroom were collected on all 32 teachers. Thus, the Classroom Arrangement treatment group had the opportunity to demonstrate competencies specific to Classroom Arrangement as well as competencies specific to Literature Experience and Sense Perception training. In addition, all of the teachers were administered a written exam by their instructors at the first and last training sessions. The exam assessed the same competencies measured by the work sample method.

The work sample data was collected by a staff of ten part-time research assistants. Trained on procedures and equipped with cameras, tape recorders, books, sensory materials, and a variety of forms and written procedures, individual research assistants made four one-day classroom visits to each of the teachers in the study sample. The first two visits took place during the two weeks immediately after training. The research assistants returned with records of the physical arrangement of each classroom and with audio-tape recordings of each teacher interacting with children in her/his own classroom during the conduct of story-reading and sense perception learning experiences. These "work samples" were rated at the Laboratory according to how well they met the criteria specified for each of the four competencies.

The FLS Teacher Education Program

Currently, the Flexible Learning System consists of an initial set of 15 learning modules for adults in early childhood education. It is

intended that these modules provide learner-centered, functional-context, competency-based experiences.

The FLS is learner-centered at three levels: at entry into the system, during the process of training, and at exit from the system. Each potential trainee who enters the system has the opportunity to select modules of personal interest. During training, teachers have the opportunity to try out various roles of persons involved in the early childhood educational environment. For example, they are provided experiences in which they play the child role, as well as experiences in which they practice teaching skills with other adults who provide feedback. When learners exit from training, they should be prepared to extend their learnings beyond the module, to "personalize" their learning to fit their individual setting, children's needs, personal, and programmatic teaching goals.

The FLS provides a functional-context experience in that it is designed for learning while on-the-job. Each module provides opportunities for daily application and practice with young children aged four to nine. The system is competency-based in that it is based on the doing of teaching acts, as well as reading and hearing about them. That is, learners participate in group discussions about child development theory and early childhood teaching, and then, as a part of their training implement their knowledge into the classroom. This cycle of discussing-practicing occurs throughout the training process.

The FLS Training Process. Although each of the 15 modules was developed in a unique style, reflecting both the nature of the subject matter and the personal style of the developer, there are six procedural steps common to the training process used in most of the modules.

First, each begins with an introduction which includes performance objectives, learning methods, and learning materials and estimates as to the length of time needed to complete the module. The second step involves a conceptual presentation of the subject matter of the module. This may involve studying the module content and participating in discussions led by the trainer. The third step is aimed at conceptual integration. Activities are presented which require personalized application of the basic concepts and ideas in the module. Once the groundwork of conceptual understanding has been laid, the fourth step requires behavioral production of particular teaching skills. This may involve role-playing sessions with other group members, or trying out the new concepts in devising lesson plans. The fifth step is classroom application. At this point, learners are provided with activity guides that enable them to apply the skills they have learned while working with children in their own classrooms. The sixth step is intended as the final outcome of each of the modules. It involves the learner's "personalization" of the training experiences for application to his/her own situation. This requires a "thinking through" of personal teaching goals and a revision and extension of FLS training to facilitate achievement of them.

Competency Definitions

The four competencies considered in this study were drawn from various FLS training modules. The quality competency was derived from the FLS module Arranging the Classroom for Children. It represented the teacher's ability to structure areas in the classroom so that they effectively serve their functions for her/him and the children. A teacher possessing this competency is able to match the instructional purpose which an area serves with the arrangement of its internal characteristics, thereby facilitating efficient use of classroom areas by the children and decreasing classroom management demands. For example, if an area is intended to facilitate development of fine motor skills through artwork activities, then it should contain appropriate artwork materials, storage of materials should encourage the child's initiation of the activity, it should be isolated from through traffic, should limit by its size the number of children able to use it, etc. To evaluate this competency, the teacher was asked a number of interview questions regarding how areas present in her/his classroom were typically used. Based on the responses, the areas present were classified according to types of areas described in the Classroom Arrangement module. In addition, each of the areas was photographed, and the resulting picture for three particular area types were rated according to quality criteria such as storage, isolation, etc. To illustrate, a teacher might indicate that an area in her classroom is designed to allow a child or two to be by themselves. The FLS module dealing with classroom arrangement defines an area with this purpose as a Private Area -- "an area with no working surfaces (e.g., desks or tables) where a child can go for privacy or isolation from everyone in the classroom." A diagram of the area is made

and photographs are taken and brought to the Laboratory to allow rating of how well the internal characteristics of the Private Area serve its purpose. These characteristics include visual isolation, no chairs, no storage of materials, protected from through traffic, and so on. A quality score was arrived at by allocating one point for each characteristic present that was recommended in the Classroom Arrangement module. Three area types were rated: a) Learning Area without Seating, b) Learning Area with Seating, and c) Private Area. Two of the areas (a and b above) had ten characteristics specified and the third area had eight. To insure equal weighting of each area, the score for the two areas with ten characteristics were multiplied by a constant of eight, while the area with eight characteristics specified was multiplied by a constant of ten. The scores for all three areas were then summed and divided by the number of areas (3) to obtain a total score.

The competency of personalization was derived from the FLS module Enriching Literature Experiences of Children. It is a key FLS competency in that it requires knowledge of the child's cultural background, her/his reactions to social situations, etc. A teacher with mastery of the competency personalization is one who incorporates the experiences of individual children into the story-reading process. The teacher demonstrated personalization if she could describe to a research assistant a child's reaction to the events during a story (e.g., "Tom knew what was going to happen next.") or by describing a child's reaction through a personal experience (e.g., "Maria remembered that the same thing that happened in the story once happened to her.") A personalization score was

computed by summing the total number of children (max=5) for which the teacher described reactions. The opportunity to demonstrate personalization was provided by the following interview questions:

1. What are the names of the children you read the story to?
2. What do you think that each of the children got out of the story-reading experience?

The competencies of diversity and awareness were derived from the FLS module Developing Children's Sense Perception. Diversity represents the teacher's skill at integrating the child's sensory and cognitive processes by using a variety of questions, statements, or explanation strategies within a sensory activity. The strategies are intended to develop the child's classification skills. The diversity score was computed by summing the number of appropriate teacher strategies used across the 5 primary senses during the sense perception experience. Each strategy was awarded one point. The awareness competency represented the ability of the teacher to identify multiple methods of using materials or objects in activities which facilitates the child's use of a variety of senses. The teacher was shown a set of pencils, then asked: "Please tell me all the ways you can think of using these pencils with the children in your class." If there was no response, she/he was asked: "What activities could you set up for children with these pencils?" When an initial list was recorded, a follow-up question was asked: "Can you think of any other possibilities?" A score was derived by summing all of the uses listed which involved children's using touch, sight, hearing, smell, or taste, multiplying this sum by 100, and dividing it by the total number of sensory plus non-sensory uses given by the teacher.

Instrumentation

Two methods, a classroom performance "work sample" rating and a paper-pencil exam, were administered to both educational subsamples to examine pre- to posttraining differences in teacher performance. Both methods paralleled one another in content subscales and were designed to measure four competencies considered critical for effective early childhood teaching. The methods represented different ends of a conceptual-behavioral performance continuum. That is, the paper-pencil method required trainees to utilize written test-taking skills while the work sample method required trainees to demonstrate competencies behaviorally or orally in a classroom situation.

Work sample method. The work sample method was designed to obtain measures of teachers' classroom application of the four competencies described. Work sample data collection began with a research assistant accompanying each teacher on a tour of his/her classroom. The assistant diagrammed the classroom's physical layout and interviewed the teacher regarding the purpose, use, and arrangement of each activity area with the room. The tour proceeded in a clockwise direction, and several photographs of each area were taken.

Following the classroom tour, each teacher conducted a story telling session with a group of 3-5 children using a children's book sent to her one week earlier. One half of the 32 teachers used the book Snowy Day on the first and third data collection occasions and the book Birthday Presents on the second and fourth occasions. The remaining 16 teachers used the books in reverse order.

The last work sample activity involved the teachers in conducting a

sensory experience using a set of beads of different colors and shapes and a set of six small containers of spices. One half of the teachers used beads for the sensory experiences on the first and third data collection occasions and spices on the second and fourth occasions. The remaining half of the teachers used the materials in reverse order: spices on first and third occasions, beads on second and fourth. The teachers conducted the experiences with a group of 3-5 children different from those involved in the story reading activities. The experiences were tape recorded and the tapes brought to the Laboratory for scoring.

Paper-pencil method. The paper-pencil exam contained four subscales, each covering a separate competency. The exam contained 16 items totalling 45 points and had no time limit specified. Table 1 shows the distribution of items and points across competencies. All items were scored at the Laboratory according to a master key. The open-ended items were scored by two raters trained to an 85% level of agreement.

The quality competency was assessed with four multiple-choice items and two short-answer items. One of the four multiple-choice items presented a classroom situation and then asked four separate questions regarding the arrangement of an area. Each of the questions was worth one point and had at least three distractors. The two short-answer items required three and four responses respectively, with each correct response receiving one point. The personalization competency was measured via three multiple-choice items, each worth one point. The diversity competency was assessed with three items. Two of the items were short-answer. One required three responses worth one point each, while the other required two responses worth one point

Table 1. Number of paper-pencil exam items and points per competency.

Competency	No. of items	No. of points
Quality	6	14
Personalization	3	3
Diversity	3	14
Awareness	4	14
Total	16	45

each. The third item required matching a hypothetical teacher's verbal statements with the particular cognitive skills of the child that each would help develop. One point was given for each correct match out of nine possible. The awareness competency was measured by four items. One items was open-ended and worth a maximum of seven points. Two of the four items were short-answer items worth two and four points respectively. The fourth items was a multiple-choice worth one point.

Results and Discussion

The multicompetency-multimethod matrices for the high and low education subsamples before and after FLS training are presented in Tables 2-5. The values in parentheses represent test-retest reliabilities for the competency subscales on the paper-pencil exam and generalizability coefficients (Cronbach, et al., 1974) for the work sample measures. The paper-pencil reliabilities were computed for each of the educational subsamples. As can be seen, the work-sample generalizability coefficients were satisfactory across the measures, but the paper-pencil reliabilities ranged between .29 and .51 for the high education subsample and .29 and .83 for the low education subsample. These low paper-pencil subscale reliabilities limit to some degree the generalizability of the findings.

Tables 6 and 7 summarize the pre- and posttraining performance on and validity of the paper-pencil and work sample methods for the high and low education subsamples. The mean heterocompetency-heteromethod values in the tables were obtained by averaging the correlations between a competency and each of the other competencies measured by a particular method.

Table 2. Pretraining multicompetency-multimethod matrix for high education teachers.

(n=15)

		Paper-Pencil				Work-Sample			
		Qual.	Person.	Div.	Awa.	Qual.	Person.	Div.	Awa.
Paper-Pencil	Quality	(.29)							
	Personalization	-.05	(.51)						
	Diversity	-.20	.01	(.29)					
	Awareness	.10	.09	.13	(.51)				
Work-Sample	Quality	(.18)	-.35	.27	.02	(.84)			
	Personalization	.11	(.13)	-.25	-.10	.48	(.67)		
	Diversity	-.36	-.28	(.09)	.16	.41	-.06	(.78)	
	Awareness	.36	.36	.08	(.15)	.33	.28	.29	(.98)

Note. The validity values are circled and the reliability values are in parentheses.

Table 3. Pretraining multicompetency-multimethod matrix for low education teachers.

(n=15)

		Paper-Pencil				Work-Sample			
		Qual.	Person.	Div.	Awa.	Qual.	Person.	Div.	Awa.
Paper-Pencil	Quality	(.37)							
	Personalization	.59	(.34)						
	Diversity	.01	-.11	(.29)					
	Awareness	.32	.09	.50	(.83)				
Work-Sample	Quality	(-.02)	.25	-.20	.04	(.84)			
	Personalization	.36	(.41)	-.03	0.03	-.10	(.67)		
	Diversity	.01	-.02	(.18)	.19	-.03	.25	(.78)	
	Awareness	-.38	-.43	.48	(.26)	.13	-.27	-.06	(.98)

Note. The validity values are circled and the reliability values are in parentheses.

Table 4. Posttraining multicompetency-multimethod matrix for high education teachers.

(n=15)

		Paper-Pencil				Work-Sample			
		Qual.	Person.	Div.	Awa.	Qual.	Person.	Div.	Awa.
Paper-Pencil	Quality	(.29)							
	Personalization	.49	(.51)						
	Diversity	-.12	-.08	(.29)					
	Awareness	.12	-.19	.07	(.51)				
Work-Sample	Quality	(.43)	.05	.30	-.02	(.84)			
	Personalization	-.19	(-.34)	-.25	-.23	-.19	(.67)		
	Diversity	-.07	-.04	(-.15)	-.01	-.10	-.27	(.78)	
	Awareness	.47	.04	.12	(.13)	.57	-.05	.27	(.98)

Note. The validity values are circled and the reliability values are in parentheses.

Table 5. Posttraining multicompetency-multimethod matrix for low education teachers.

(n=15)

		Paper-Pencil				Work-Sample			
		Qual.	Person.	Div.	Awa.	Qual.	Person.	Div.	Awa.
Paper-Pencil	Quality	(.37)							
	Personalization	-.02	(.34)						
	Diversity	-.17	.44	(.29)					
	Awareness	.44	.25	.52	(.83)				
Work-Sample	Quality	(.70)	.09	.04	.25	(.84)			
	Personalization	-.30	(.00)	.18	.13	-.11	(.67)		
	Diversity	.59	.06	(.20)	.60	.48	.32	(.78)	
	Awareness	-.06	-.14	.01	(.36)	-.13	.30	-.01	(.98)

Note. The validity values are circled and the reliability values are in parentheses.

Convergent validity. Table 6 indicates that moderate confirmation of the competencies by independent measurement procedures was achieved before training for both the high and low education subsamples. A major exception occurred in the low education sample on the quality competency, for which there was no indication of convergent validity. With this one exception, the competencies seem to be reflecting the teaching constructs they are intended to in the sense that they are confirmed by independent measurement procedures (Campbell and Fiske, p. 83). The picture changes after training, however (Table 7). The mean validity value across competencies decreased for the high education sample but increased for the low education sample. Thus, the competencies seem to become better defined or identified as a result of training for the low but not the high education subsample. This trend did not hold, however, for the personalization competency. Its validity decreased substantially after training for both samples. The validity value for quality, on the other hand, increased substantially after training for both samples, suggesting that the competency became very well defined during the training process.

Discriminant validity. A second aspect of validity is how well the measures can distinguish the competencies from one another. Tests can be invalidated because of too high correlations with other tests purporting to measure different things (Campbell and Fiske, 1959, p. 84). This type of invalidation occurs when heterocompetency-heteromethod values are as high as or exceed the validity values. This does not seem to be a problem in this study however. For both samples, the mean heterocompetency-heteromethod values were consistently lower than the validity values both before and after training, suggesting that the competencies are independent of one

Table 6. Pretraining means and competency interccorrelations for high and low education teachers.

	High Education (n=16)						Low Education (n=15)					
Col. No.	1	2	3	4	5	6	7	8	9	10	11	12
	Mean Score											
	PP	WS	Val.	Mean Het-Het	Mean Hetero-PP	Mean Hetero-WS	PP	WS	Val.	Mean Het-Het	Mean Hetero-PP	Mean Hetero-WS
Quality	7.3	46.4	.18	.008	-.05	.41	6.5	47.0	-.02	.01	.31	.00
Personalization	1.7	0.3	.13	.09	-.03	.23	1.6	0.6	.41	.10	.19	-.04
Diversity	9.9	6.3	.09	-.06	-.02	.21	7.6	6.9	.18	.008	.13	.05
Awareness	5.0	4.7	.15	.11	.30	.29	3.9	6.0	.26	-.09	.30	-.07
$Mean_c$			.14	.002	.003	.29			.21	.007	.24	-.00

Note. Col. No. = column number; Val. = validity values; Mean Het-Het = average heterocompetency-heteromethod value; Mean Hetero-PP = average paper-pencil heterocompetency-monomethod value; Mean Hetero-WS = average work-sample heterocompetency-monomethod value; $Mean_c$ = average value per column.

Table 7. Posttraining means and competency intercorrelations for high and low education teachers.

	High Education (n=16)						Low Education (n=15)					
Col. No.	1	2	3	4	5	6	7	8	9	10	11	12
	Mean Score											
	PP	WS	Val.	Mean Het-Het	Mean Hetero-PP	Mean Hetero-WS	PP	WS	Val.	Mean Het-Het	Mean Hetero-PP	Mean Hetero-WS
Quality	8.17	49.8	.43	.09	.16	.09	7.2	49.8	.70	.10	.08	.08
Personalization	1.67	0.58	-.34	-.10	.07	-.17	2.1	0.82	.00	.003	.22	.17
Diversity	11.6	6.3	-.15	.008	-.04	-.03	10.1	7.5	.20	.25	-.08	.26
Awareness	5.5	5.2	.13	.06	-.05	.07	5.3	7.9	.36	.13	.40	.05
$Mean_c$			.02	.02	.04	-.01			.32	.12	.23	.14

Note. Col. No. = column number; Val. = validity values; Mean Het-Het = average heterocompetency=heteromethod value; Mean Hetero-PP = average paper-pencil heterocompetency-monomethod value; Mean Hetero-WS = average work-sample heterocompetency-monomethod value; $Mean_c$ = average value per column.

another.

Another test of discriminant validation is whether "... a variable correlate higher with an independent effort to measure the same trait than with measures designed to get at different traits which happen to employ the same method (Campbell and Fiske, 1959, p.83)." Essentially, this means that the validity coefficients be larger than the correlations in the monomethod triangles, indicating that the method accounts for minimal score variance.

An analysis of method variance. This aspect of validity is a critical issue in this paper, as there is reason to suspect that a method may influence demonstration of discriminable competencies. Inspection of Table 6 reveals a striking difference in which method accounted for score variance among the subsamples before training. For the high education subsample, the low heterocompetency values indicate that the paper-pencil measure has no method factors (e.g., test form factors, response sets) which generalize across the competencies. The pattern changes for the work-sample method, however. The correlations among competencies measured with this method surpass the validity values as well as the heterocompetency-heteromethod values. The implication is that the work-sample method employs procedures which limit the high education sample's demonstration of the competencies. Interestingly, however, the pattern of method variance was reversed in the low education subsample. The correlations among competencies measured with the work sample method were low, indicating that the method had minimal influence on the demonstration of each of the competencies. The correlations between the competencies when measured with the paper-pencil method were

uniformly higher than the validity values and heterocompetency-heteromethod values. The paper-pencil test seems to have features which generalize across the competencies and reduce their discriminant validity.

What happens, however, when educational training is administered? Does the training minimize amount of variance accounted for by a method? This seems to be true in a relative sense for the high education sample. The variance contributed by either method is minimal, suggesting that there is no longer an "apparatus" or procedures factor which accounts for their score variance. Note, however, that the convergent validity drops for every competency except quality, which shows a substantial increase. The low correlations across the columns in Table 6 for the high education sample seem to suggest the competencies become fragmented, with each method measuring separate aspects of each particular competency. It seems as if the high education teachers tend to dissociate the "showing" (performance via the work-sample method) a competency from "telling" (performance via the paper-pencil exam) about it. Their consistently higher mean scores for each of the competencies after training suggests that even if such dissociation does occur, the high education teachers can nonetheless demonstrate improved performance.

The mean scores for the low education teachers also increase following training. In addition, convergent validity improves to the extent that the average paper-pencil method variance no longer surpasses it. It is interesting to note however, that the average correlations between competencies measured with the paper-pencil exam still tend to be high, suggesting that method factors still exist but are not strong enough to prevent demonstration of discriminable competencies.

It should be noted here that most of what has been said above has been based on general trends or means. A close inspection of each competency validity would yield valuable information regarding the interaction of competencies and methods with educational samples. For example, personalization decreases substantially in validity for both samples after training. This finding suggests that the construct was not adequately operationalized by either method, especially since the pre and post mean work sample scores were low for both groups and there was only a small number of personalization items on the paper-pencil test. Quality, on the other hand, showed an encouraging increase in validity and mean performance scores for both samples after training. The method variance contributed by the work sample method for the high education teachers and the paper-pencil exam for the low education teachers decreased substantially, suggesting that both samples learned how to show and tell about the internal arrangement of classroom areas.

Summary

The principle issue addressed in this study consists of the appropriateness of the test contructor's constructs of competencies and methods for measuring them with teachers differing in educational background. When a construct is operationalized in the form of a measure, care is usually exercised in the development of test form factors such as item content and format, administration, and instructions. The careful test user usually examines the quality of these factors and then uses the test to determine the match of the examinee's construct of a competency with the test

constructor's. With the limits of low paper-pencil subscale reliabilities in mind, this study suggests a weakness in this process -- namely, the appropriateness of the entire method for the persons it will be used with. Pretraining assessment of high and low education teachers indicated that particular methods tended to contain procedures which generalized across competencies and to a degree invalidated them as dintinguishable indicators of constructs. For the high education sample, the work-sample method accounted for variance in performance scores. Perhaps teachers with a relatively high amount of education have difficulty demonstrating competencies in the classroom in the manner required by the work-sample method, but have no such difficulty demonstrating them on the paper-pencil exam -- a method with which their educational experience makes them highly familiar.

On the other hand, the teachers with relatively less education tend to have minimal score variance accounted for by the work-sample method before training, suggesting that they can demonstrate competencies in the classroom situation. They are not so fortunate with the paper-pencil method. It appears to soak up some of their score variance, suggesting that the test has generalizable factors such as item format, writing requirements, etc. which extend across competencies and limit their differentiation from one another.

In general, these differences in method variance among educational subsamples tended to disappear following training. It is important to note that the most likely reason for this is that the training was based on a process linking theory of child development with an experiential approach to early childhood teaching. Had the link not been made, perhaps the method variance would not have decreased for either sample.

References

Alward, K. R. Arranging the Classroom for Children. San Francisco: Far West Laboratory for Educational Research and Development, 1973.

Campbell, D. T., & Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 1959, 56, 2, 81-105.

Cronbach, L. J., Gleser, G. C., et al. The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles. New York: John Wiley and Sons, Inc., 1972.

Gilbert, B. L. & Finzer, W. F. Developing Children's Sense Perception. San Francisco: Far West Laboratory for Educational Research and Development, 1975.

Taylor, J. Enriching Literature Experiences of Children. San Francisco: Far West Laboratory for Educational Research and Development, 1975.

Taylor, T. D., Hubner, J. J., Chow, S. H. L., Stripp, S. P., & Tanaka, M. N. Final Report on Evaluation and Research of the Flexible Learning System: an Early Childhood Teacher Education Program. San Francisco: Far West Laboratory for Educational Research and Development, 1975.

Taylor, T. D., Chow, S. H. L. Main Field Test Report for Five Flexible Learning System Competency-Based Training Modules - Milestone B, Work Unit IV. San Francisco: Far West Laboratory for Educational Research and Development, 1974.